Estadística multivariada, 1 sem. 2019

Juan Carlos Castillo & Alejandro Plaza

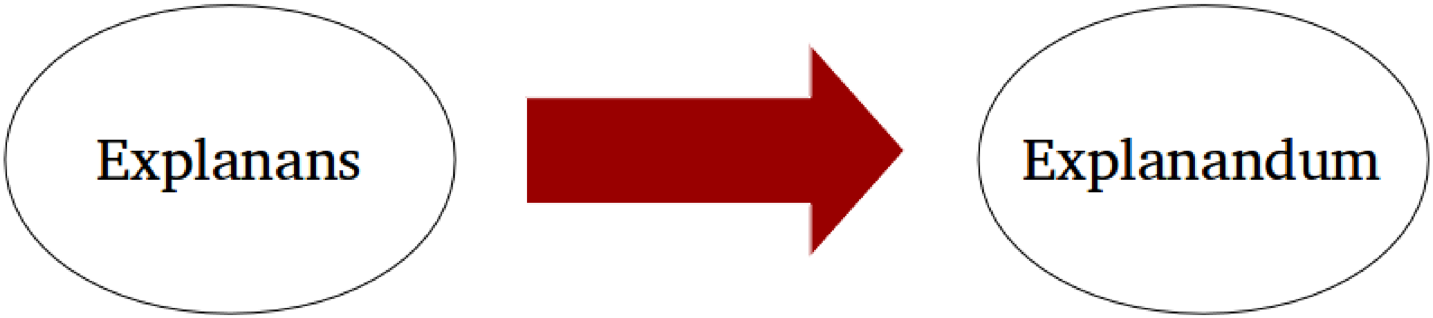
Sesión 3: Regresión simple 1

Contenidos

1. Repaso de sesión anterior
2. Regresión simple
3. Actividad práctica

1. Repaso sesión anterior

El concepto de explicación en ciencias sociales



- Explanandum: el fenómeno que predentemos explicar (precisión, relevancia y variabilidad).
- Explanans: lo que genera la aparición del fenómeno (lógica, eficacia y claridad.)

Dispersión: Varianza

- Suma de las diferencias al cuadrado de cada valor (x) y el promedio de la distribución divididos por el total menos 1. Formalmente:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

- Considerando N-1 para la varianza de la muestra.

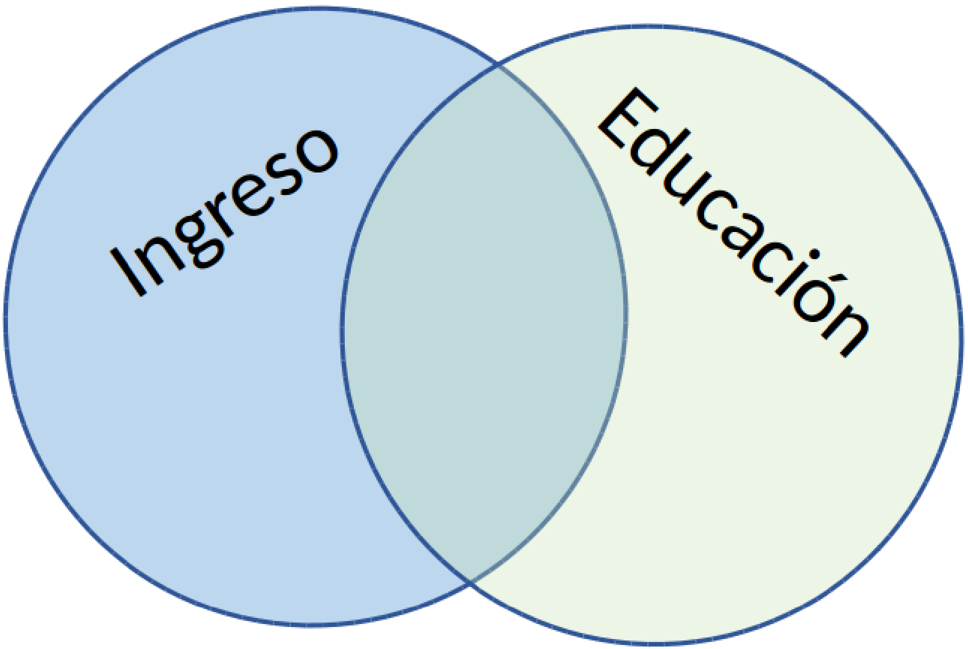
ID	**Pje (x)**	$x - \bar{x}$	$(x - \bar{x})^2$
1	6	0.4	0.16
2	4	-1.6	2.56
3	7	1.4	1.96
4	2	-3.6	12.96
5	9	3.4	11.56
Sum	28	0	29.2
Prom	5.6		

$$\sigma^2 = \frac{(29.2)}{5-1}$$

$$= 7.3$$

Asociación: covarianza / correlación

¿Se relaciona la variación de una variable, con la variación de otra variable?



Asociación: covarianza / correlación (II)

- Covarianza

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlación

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}$$

O bien

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Ejemplo de correlación

Estimar la correlación entre puntaje en lenguaje (x) y puntaje en matemáticas (y):

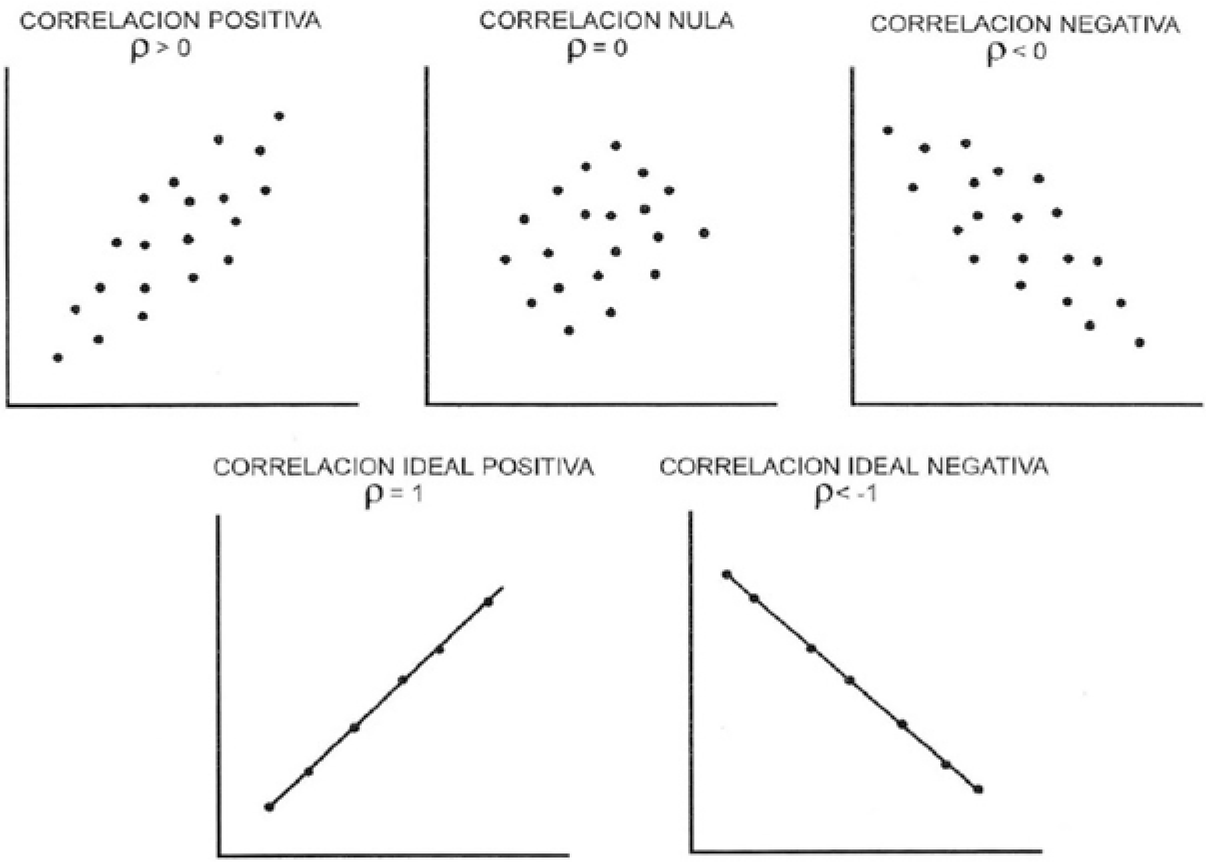
id	x	y	(A) $x-\bar{x}$	(B) $y-\bar{y}$	A*B	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	17	24	-3	3	-9	9	9
2	19	23	-1	2	-2	1	4
3	14	22	-6	1	-6	36	1
4	22	17	2	-4	-8	4	16
5	15	23	-5	2	-10	25	4
6	26	21	6	0	0	36	0
7	23	18	3	-3	-9	9	9
8	21	17	1	-4	-4	1	16
9	28	21	8	0	0	64	0
10	15	24	-5	3	-15	25	9
Sum					-63	210	68
Prom	20	21					

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$= \frac{-63}{\sqrt{210 * 68}}$$

$$= -0.5272$$

Nube de puntos (scatterplot) y correlación



¿Preguntas?

2. Modelo de regresión simple

Objetivos centrales del modelo de regresión:

1. Conocer la variación de una variable (dependiente, Y) de acuerdo a la variación valor de otra variable (independiente, X):

   - *Ej: En qué medida el puntaje PSU influye en el éxito académico en la universidad?*

2. Estimar el valor de una variable de acuerdo al valor de otra (predicción)

   - *Ej: Si una persona obtiene 600 puntos en la PSU, que promedio de notas en la universidad es probable que obtenga? (Atención: predicción no implica explicación)*

3. Establecer en que medida esta asociación es significativa (inferencia)

   - *¿Se puede generalizar a la población? ¿Con qué nivel de confianza?*

Terminología

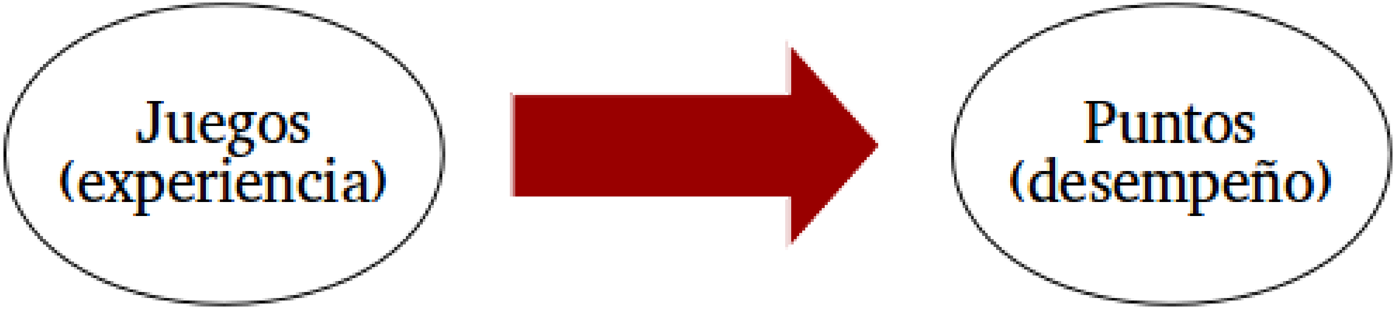
TABLA 2.1

Terminología en la regresión simple

y	x
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictora
Regresando	Regresor

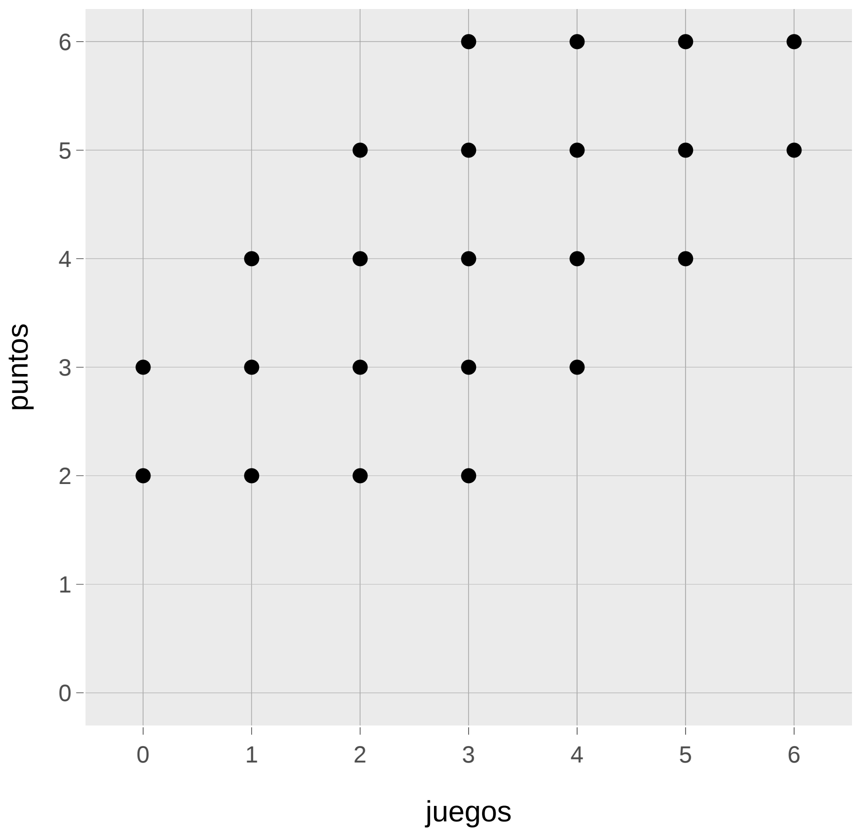
Ejemplo

¿En qué medida la experiencia previa jugando un juego predice el nivel de puntos (en juego posterior)?



Datos

	id	juegos	puntos
1	1	0	2
2	2	0	3
3	3	1	2
4	4	1	3
5	5	1	4
6	6	2	2
7	7	2	3
8	8	2	4
9	9	2	5
10	10	3	2
11	11	3	3
12	12	3	4
13	13	3	5
14	14	3	6
15	15	4	3
16	16	4	4
17	17	4	5
18	18	4	6
19	19	5	4
20	20	5	5
21	21	5	6
22	22	6	5
23	23	6	6



Descriptivos

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
id	23	12.000	6.782	1	6.5	17.5	23
juegos	23	3.000	1.758	0	2	4	6
puntos	23	4.000	1.382	2	3	5	6

Idea de distribución condicional

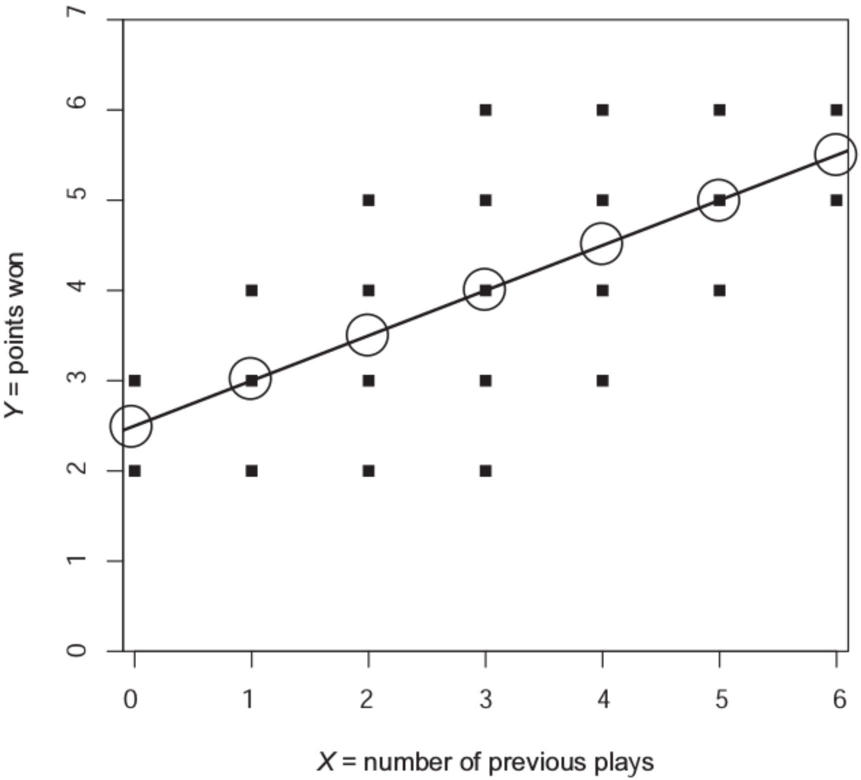
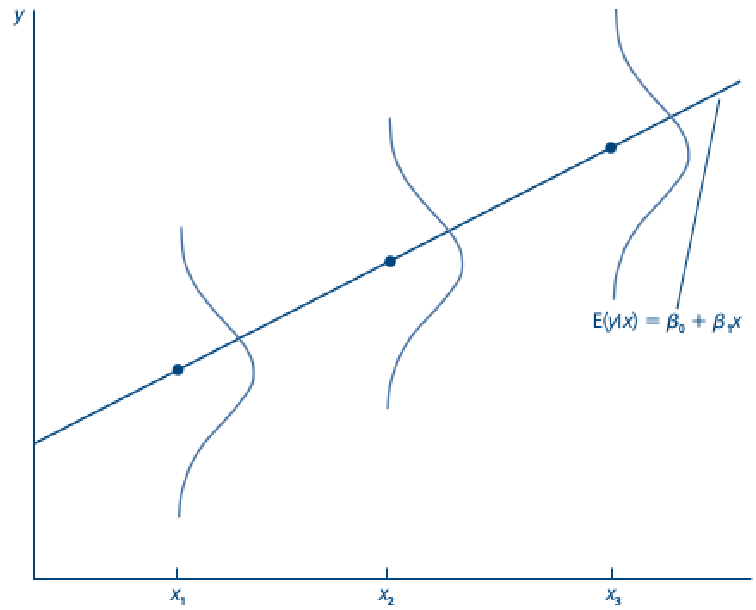


FIGURE 2.2. A line through conditional means.

Idea de distribución condicional

FIGURA 2.1

E(*y*|*x*) como función lineal de *x*.

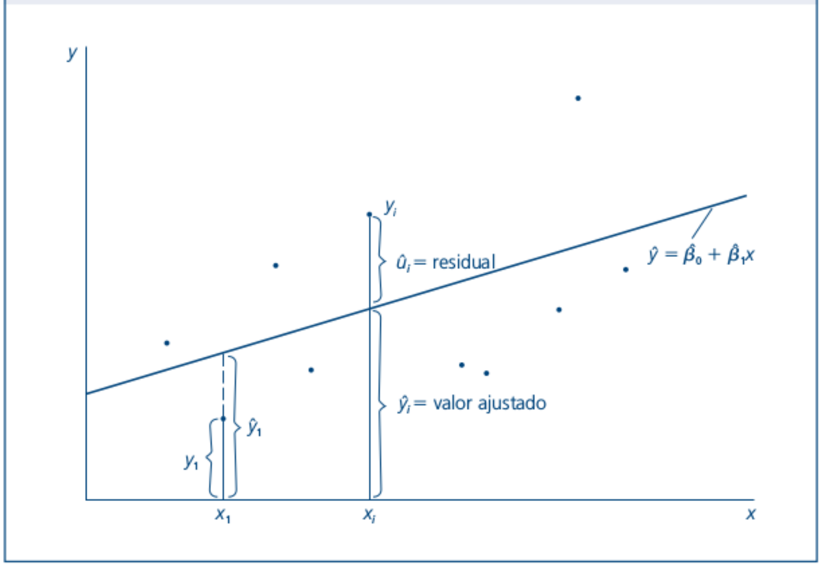


La recta de regresión

La (co) variación general de Y respecto a X se puede expresar en una ecuación de la recta = modelo de regresión

FIGURA 2.4

Valores ajustados y residuales.



Para obtener la “mejor recta” se utiliza la estimación de mínimos cuadrados (EMC, o **OLS** – Ordinary Least Squares), que minimiza la suma de los cuadrados de las distancias entre las observaciones y la recta en el eje vertical

Componentes de la ecuación de la recta de regresión

$$\widehat{Y} = b_0 + b_1 X$$

Donde

- $\widehat{Y}$ es el valor estimado de Y
- b_0 es el intercepto de la recta (el valor de Y cuando X es 0)
- b_1 es el coeficiente de regresión, que nos dice cuánto aumenta Y por cada punto que aumenta X

Estimación de los coeficientes de la ecuación:

$$b_1 = \frac{Cov(XY)}{VarX}$$

$$b_1 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}}$$

Y simplificando

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}$$

Luego despejando el valor de b_0

$$b_0 = \bar{Y} - b_1\bar{X}$$

Cálculo basado en el ejemplo

la base para todos estos calculos es la diferencia de cada valor menos su promedio. Vamos a crear un vector en nuestra base de datos $difx = x - \bar{x}$ y $dify = y - \bar{y}$

```
datos$difx=datos$juegos-mean(datos$juegos)
datos$dify=datos$puntos-mean(datos$puntos)
```

Y ahora con esto podemos obtener la diferencia de productos cruzados $dif_cru = (x - \bar{x}) * (y - \bar{y})$, así como la suma de cuadrados de X $SSx = (x - \bar{x})^2$

```
datos$dif_cru=datos$difx*datos$dify
datos$SSx=datos$difx^2
```

Datos y vectores (columnas) adicionales

```
datos
```

```
##     id juegos puntos difx dify dif_cru SSx
## 1    1      0      2   -3   -2       6   9
## 2    2      0      3   -3   -1       3   9
## 3    3      1      2   -2   -2       4   4
## 4    4      1      3   -2   -1       2   4
## 5    5      1      4   -2    0       0   4
## 6    6      2      2   -1   -2       2   1
## 7    7      2      3   -1   -1       1   1
## 8    8      2      4   -1    0       0   1
## 9    9      2      5   -1    1      -1   1
## 10  10      3      2    0   -2       0   0
## 11  11      3      3    0   -1       0   0
## 12  12      3      4    0    0       0   0
## 13  13      3      5    0    1       0   0
## 14  14      3      6    0    2       0   0
## 15  15      4      3    1   -1      -1   1
## 16  16      4      4    1    0       0   1
## 17  17      4      5    1    1       1   1
## 18  18      4      6    1    2       2   1
## 19  19      5      4    2    0       0   4
## 20  20      5      5    2    1       2   4
## 21  21      5      6    2    2       4   4
## 22  22      6      5    3    1       3   9
## 23  23      6      6    3    2       6   9
```

Cálculo basado en el ejemplo

Y con esto podemos obtener la suma de productos cruzados y la suma de cuadrados de X

```
sum(datos$dif_cru)
```

```
## [1] 34
```

```
sum(datos$SSx)
```

```
## [1] 68
```

Reemplazando en la fórmula

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})} = \frac{34}{68} = 0.5$$

Cálculo basado en el ejemplo

Reemplazando podemos obtener el valor de b_0

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 4 - (3 * 0.5) = 2.5$$

Completando la ecuación:

$$\widehat{Y} = 2.5 + 0.5X$$

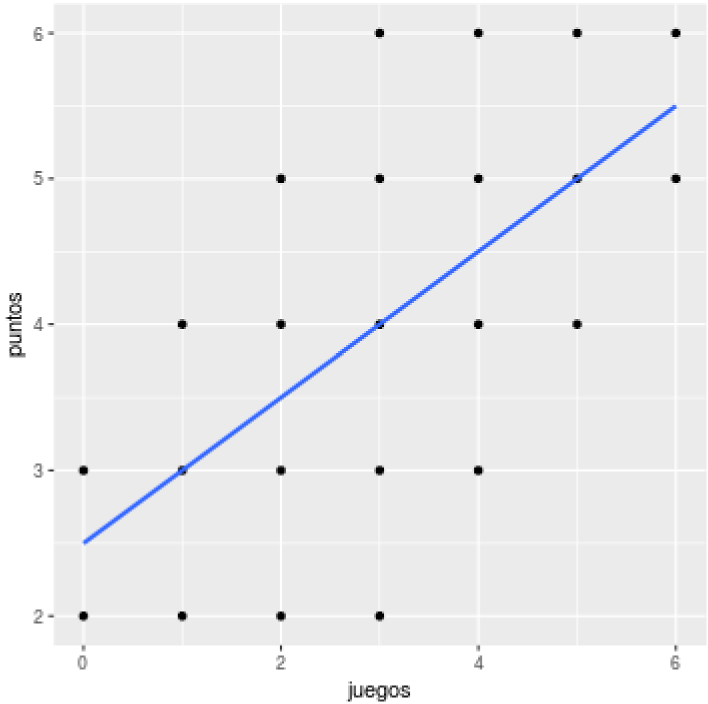
Esto nos permite estimar el valor de Y (o su media condicional) basado en el puntaje X. Por ejemplo, cuál es el valor estimado de Y dado $X = 3$?

$$\widehat{Y} = 2.5 + (0.5 * 3)$$

$$\widehat{Y} = 2.5 + (0.5 * 3) = 4$$

Cálculo basado en el ejemplo

```
ggplot(datos, aes(x=juegos, y=puntos)) + geom_point() +
geom_smooth(method=lm, se=FALSE)
```



Regresión simple en R

Estimación del modelo de regresión simple en R

La función para estimar regresión en `R` es `lm` (linear model). Su forma general es:

```
objeto=lm(dependiente ~ independiente, data=datos)
```

Donde

- objeto: el nombre (cualquiera) que le damos al objeto donde se guardan los resultados de la estimación
- dependiente / independiente: los nombres de las variables en los datos
- data = el nombre del objeto de nuestros datos en R

Estimación del modelo de regresión simple en R

En nuestro ejemplo:

```
reg1 <-lm(puntos ~juegos, data = datos)
```

`reg1` es el objeto que almacena la información de nuestra estimación. Para un reporte simple:

```
reg1
```

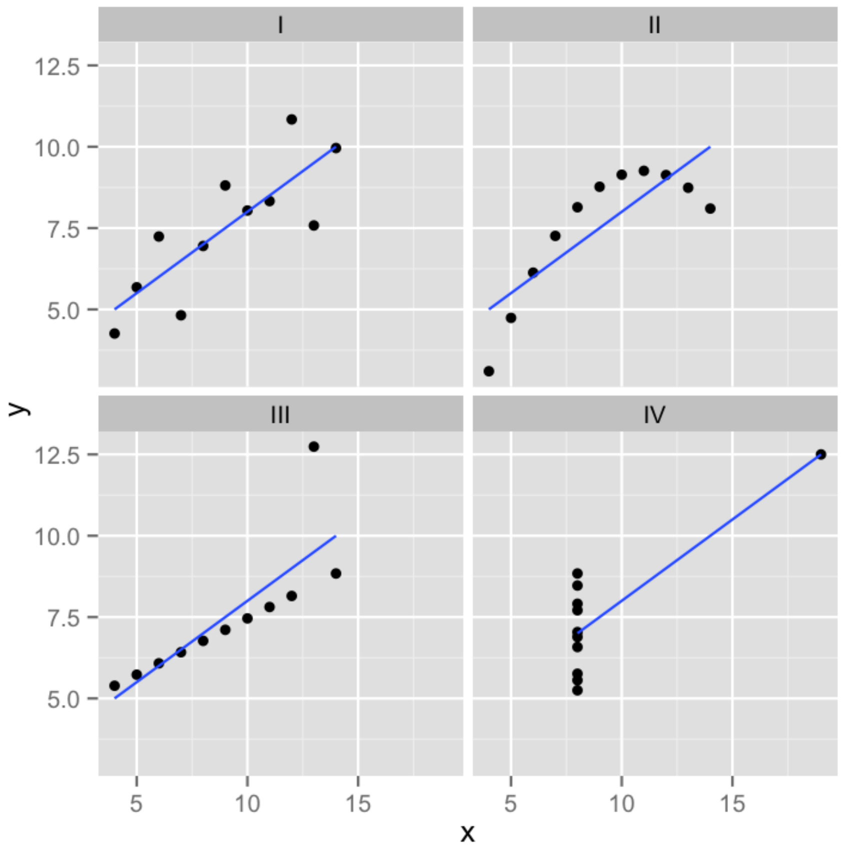
```
##
## Call:
## lm(formula = puntos ~ juegos, data = datos)
##
## Coefficients:
## (Intercept)       juegos
##         2.5          0.5
```

Y en formato más publicable

```r
stargazer(reg1, type = "html")
```

	Dependent variable:
	puntos
juegos	0.500^{***}
	(0.132)
Constant	2.500^{***}
	(0.458)
Observations	23
R^2	0.405
Adjusted R^2	0.376
Residual Std. Error	1.091 (df = 21)
F Statistic	14.280^{***} (df = 1; 21)
Note:	*p<0.1; **p<0.05; ***p<0.01

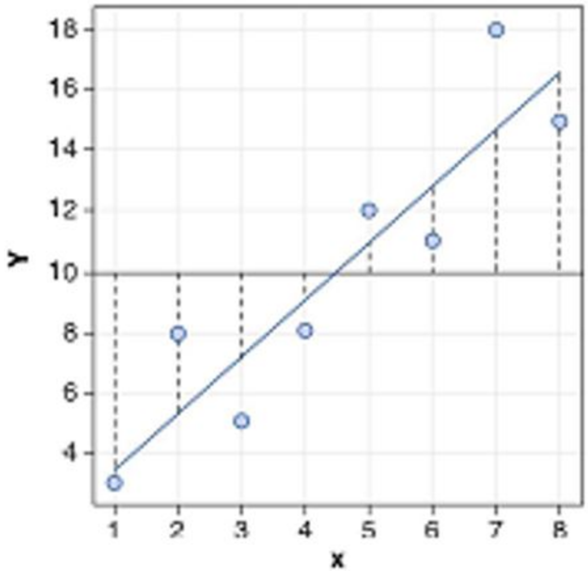
Excurso: El cuarteto de Anscombe (1973)



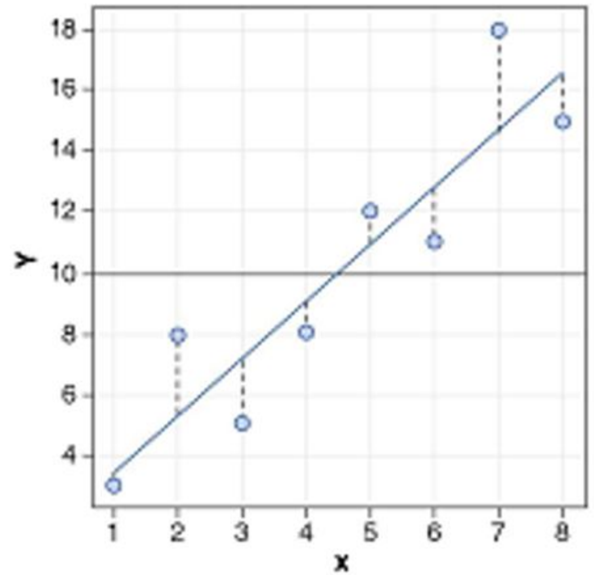
Descomponiendo Y

- Tres piezas de información relevante:
  - Valor observado de Y
  - Estimación de Y a partir de X =(Y')
  - Promedio de Y: ($\bar{Y}$)

Descomponiendo Y



$(Y' - \bar{Y})$



$(Y - Y')$

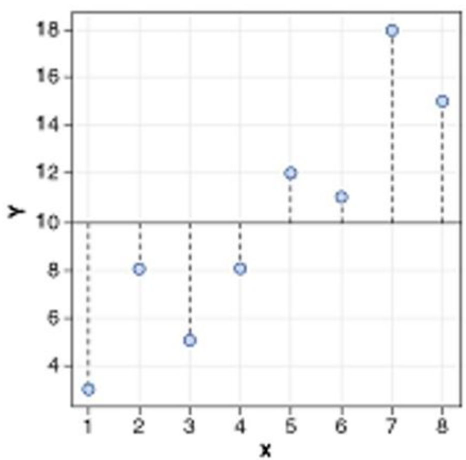
$$Y = \bar{Y} + (Y' - \bar{Y}) + (Y - Y')$$

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\bar{y} - \hat{y}_i)^2 + \Sigma(y_i - \hat{y}_i)^2$$

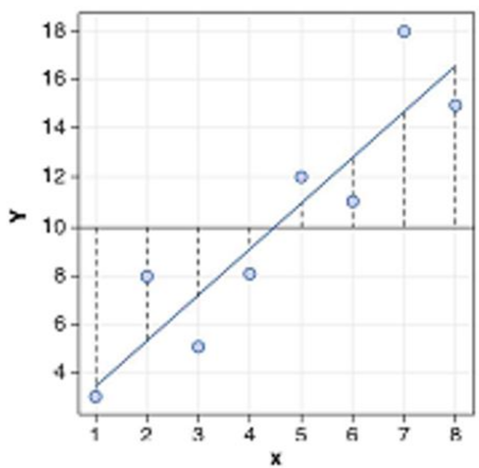
Descomponiendo Y

Conceptualmente:

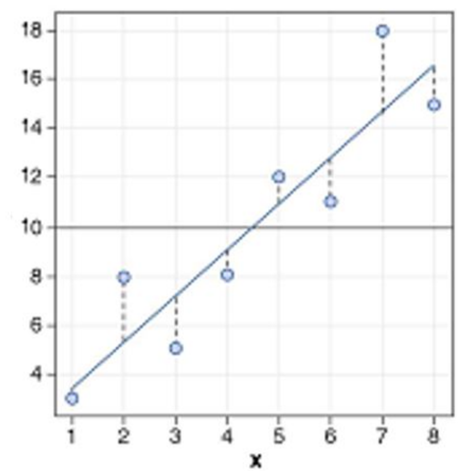
$$SS_{tot} = SS_{reg} + SS_{error}$$



SS_{tot}:
Diferencias entre valor observado de Y y el promedio de Y



SS_{reg}:
Diferencias entre valor estimado de Y y el promedio de Y



SS_{error}:
Diferencias entre valor observado de Y y la línea de regresión (valor estimado)

Descomponiendo Y

Por lo tanto:

$$SS_{tot} = SS_{reg} + SS_{error}$$

$$\frac{SS_{tot}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}} + \frac{SS_{error}}{SS_{tot}}$$

$$1 = \frac{SS_{reg}}{SS_{tot}} + \frac{SS_{error}}{SS_{tot}}$$

$$\frac{SS_{reg}}{SS_{tot}} = R^2$$

Estadística multivariada, 1 sem. 2019

Juan Carlos Castillo & Alejandro Plaza

Sesión 3: Regresión simple 1